

Data Understanding and Preparation

Introduction

Explain which original datasets we have, attributes description, number of variables, number of observations. . .

Data Exploratory Analysis

Maybe some plots here about the distribution of some of the variables. . . we'll see. I suggest to follow the same points than in the homework scripts.

Preprocessing

Loading the datasets.

```
# Load individual datasets
white <- read_delim("../data/raw/winequality-white.csv", ";", escape_double = FALSE, trim_ws = TRUE)
red <- read_delim("../data/raw/winequality-red.csv", ";", escape_double = FALSE, trim_ws = TRUE)
```

Prepare the dataset in order to create a new dataset

```
# Create a new column, where 0 indicates white wine and 1 red wine.
white$type = 0
red$type = 1

# Combine both datasets.
df <- rbind(white, red)

# Rename columns.
colnames(df) <- c("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar",
                  "chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide", "density",
                  "pH", "sulphates", "alcohol", "quality", "type")

# Quality and type are categorical variables.

df$quality <- as.factor(df$quality)
df$type <- as.factor(df$type)
```

Remove missing values and shuffle the dataset.

```
# Remove missing values.
df <- df[complete.cases(df), ] # two entries.

# Shuffle dataset.
df <- df[sample(nrow(df)),]

# Write dataset into csv file.
write.table(df, file="../data/processed/wines.csv", sep=";", col.names = TRUE, row.names = FALSE)
```